Kiejstut Bunikiewicz

Professor Brahma

BSAN 6070 – Machine Learning

2/22/21

CA03 – Decision Tree Algorithm: Question Responses

Q.1.1 Why does it make sense to discretize columns for this problem?

For this problem it makes sense to discretize the columns because some of the columns being measured like age, hours per week, and education consist of variables that are continuous. Discretizing the columns for this problem allows us to effectively group the rows of data, so we can more easily run analysis.

Q.1.2 What might be the issues (if any) if we DID NOT discretize the columns?

If we did not discretize the columns, we might have issues with speed in the program because there are over 48,000 rows of data being used. If we do not discretize the data, it may become difficult to classify the data because the continuous values may lead to an extremely large number of groups and may not lead to effective classification insights and prediction for real world data.

Q.7.1 Decision Tree Hyper-Parameter Variation vs. Performance

For ease of reading, I have added the performance metrics with only 3 digits after the decimal place.



Q.8.1 How long was your total run time to train the model?

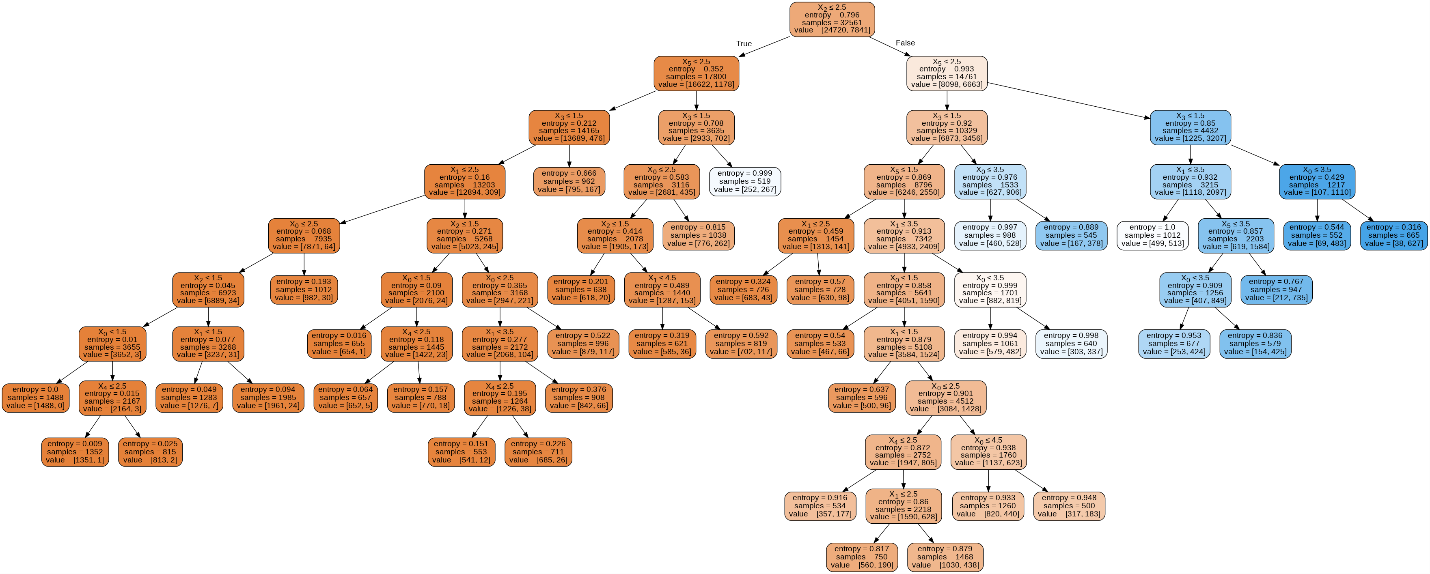
Using wall time as a metric to train my model, the total time to manually train my model was 48.9µs. The individual times ranged from 5.01µs to 7.39µs for each iteration (of 8 iterations) of the model I ran.

Q.8.2 Did you find the BEST TREE?

I do not know if I found the absolute best tree. However, I did find a tree that had a better performance than the others based on the F1 Score result. This tree used the entropy criterion, had a minimum sample leaf of 500, and a maximum depth of 10. It is possible that a combination of values from the testing could create an even more effective tree.

Q.8.3 Draw the Graph of the BEST TREE using GraphViz.

I have inserted an image of the best tree I made using GraphViz. The tree is also available in the Python notebook.



Q.8.4 What makes it the best tree?

This model is the best tree I have because it has the highest F1 Score of all of the models I produced. The F1 Score is the Harmonic Mean of Precision and Recall. This means the F1 score indicates the model with the best measure of both Precision (classifying positives as correctly as possible) and Recall (classifying negatives as correctly as possible). Precision and Recall are inversely related, so having a high F1 Score means this model captures both positives and negatives most effectively, while other models may capture one or the other more effectively than this model.

Q.10.1 What is the probability that your prediction for this person is accurate?

The prediction for the person is that they do have an income above $50,000. The probability that this prediction is accurate is a 94.28 % probability, according to the probability prediction function. The overall accuracy of the model for the entire test dataset was approximately 88.3% accuracy.